DOCUMENT RESUME

ED 216 874 SE 037 439

AUTHOR O'Brien, Francis J., Jr.
TITLE Proof that the Sample Bivariate Correlation Coefficient Has Limits ((Plus or Minus)) 1.
PUB DATE 82
NOTE 32p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Correlation; *Educational Research; Higher Education; Mathematical Formulas; *Proof (Mathematics); *Research Tools; *Statistics
IDENTIFIERS *Applied Statistics

ABSTRACT

This paper presents in detail a proof of the limits of the sample bivariate correlation coefficient which requires only knowledge of algebra. Notation and basic formulas of standard (z) scores, bivariate correlation formulas in unstandardized and standardized form, and algebraic inequalities are reviewed first, since they are necessary for understanding the proof. Then the proof is presented, with an appendix containing additional proofs of related material. (MNS)

Proof That the Sample Bivariate Correlation Coefficient Has Limits ± 1

Francis J. O'Brien, Jr., Ph.D.

Virtually all social science students who have studied applied statistics have been introduced to the concepts and formulas for linear correlation of two variables. Applied statistics textbooks routinely report the theoretical limits of the bivariate correlation coefficient; namely, that the coefficient is no more than +1 and no less than -1. However, no commonly used applied statistics textbook proves this. One of the best textbooks available to students of education and psychology introduces the proof (Glass and Stanley, 1970). Undoubtedly, one of the constraints placed on authors by publishers is space limitations available for detailed explanations, derivations and proofs.

This paper will set forth in detail a proof of the limits of the sample bivariate correlation coefficient. Since the proof requires only knowledge of algebra, most students of applied statistics at the advanced undergraduate or introductory graduate level should have little difficulty in understanding the proof. As a former instructor of graduate level introductory applied statistics, I know that the typical student can understand the proof as it is presented here.

The key for understanding statistical proofs is a presentation of detailed steps in a well articulated and coherent manner. A review of relevant statistical and mathematical concepts is also helpful (and usually required). When students are presented in detail important statistical proofs, they feel that some of the mystery and magic of mathematics has been unveiled. My experience has been that the typical student of applied statistics can follow a good number of proofs because most proofs can be presented algebraically without use of calculus. In addition to enhancing knowledge, an occasional proof often increases academic motivation.[1]

Some Preliminary Concepts

The proof requires knowledge of several concepts in statistics and mathematics. In order to make this paper self-contained, some preliminary concepts stated in a consistent notation will be reviewed. We will review the concepts and formulas of standard scores (z scores), bivariate correlation formulas in unstandardized and standardized form, and algebraic inequalities.

Notation and Basic Formulas

Table 1 is a layout of symbolic values written in the notation to be used in this paper. The model presented in Table 1 is of two measures in unstandardized (raw score) and standardized (z score) form. Table 2 presents some familiar formulas based on unstandardized variables that will be useful for the development of the proof.

[1] This paper is one of a series contemplated for publication. [See O'Brien, 1982]. Eventually I hope to present a textbook of applied statistics proofs and derivations to supplement standard applied statistics textbooks.

Table 1

Table Layout for Two Measures in Unstandardized and Standardized Form

	Measure X		Measure Y	
	Unstandardized	Standardized	Unstandardized	Standardized
	X_1	$(X_1-\bar{X})/S_x = Z_{x_1}$	Y_1	$(Y_1-\bar{Y})/S_y = Z_{y_1}$
	X_2	$(X_2-\bar{X})/S_x = Z_{x_2}$	Y_2	$(Y_2-\bar{Y})/S_y = Z_{y_2}$
	X_3	$(X_3-\bar{X})/S_x = Z_{x_3}$	Y_3	$(Y_3-\bar{Y})/S_y = Z_{y_3}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	X_i	$(X_i-\bar{X})/S_x = Z_{x_i}$	Y_i	$(Y_i-\bar{Y})/S_y = Z_{y_i}$
	.	.	.	.
	.	.	.	.
	.	.	.	.
	X_n	$(X_n-\bar{X})/S_x = Z_{x_n}$	Y_n	$(Y_n-\bar{Y})/S_y = Z_{y_n}$
Sample Size	n_x	n_{z_x}	n_y	n_{z_y}
Sample Mean	$\bar{X}$	$\bar{Z}_x$	$\bar{Y}$	$\bar{Z}_y$
Sample Variance	S_x^2	$S_{z_x}^2$	S_y^2	$S_{z_y}^2$

NOTE: all sample size terms are equal; that is: $n_x = n_{z_x} = n_y = n_{z_y}$

Any of these sample size terms could be identified by just one symbol -- such as n. We will use n when it is not important to distinguish among the other sample size terms, but will use the table values above when it is necessary or important to do so.

Table 2

Relevant Formulas for Unstandardized Measures

	Measure X	Measure Y
Sample Mean	$\bar{X} = \frac{\sum_{i=1}^{n_x} X_i}{n_x}$	$\bar{Y} = \frac{\sum_{i=1}^{n_y} Y_i}{n_y}$
Sum	$n_x\bar{X} = \sum_{i=1}^{n_x} X_i$	$n_y\bar{X} = \sum_{i=1}^{n_y} Y_i$
Sample Variance	$S_x^2 = \frac{\sum_{i=1}^{n_x} (X_i-\bar{X})^2}{n_x-1}$	$S_y^2 = \frac{\sum_{i=1}^{n_y} (Y_i-\bar{Y})^2}{n_y-1}$
Sum of Squares	$(n_x-1)S_x^2 = \sum_{i=1}^{n_x} (X_i-\bar{X})^2$	$(n_y-1)S_y^2 = \sum_{i=1}^{n_y} (Y_i-\bar{Y})^2$

NOTES:

1. The sample size terms are equal: $n_x=n_y$. Also, $n_x=n_y=n$.

2. "Sum"is simply an algebraic manipulation of "Sample Mean"; i.e., multiply over the sample size term in "Sample Mean" to get "Sum". Also,"Sum of Squares" is such a manipulation based on "Sample Variance".. "Sum" and "Sum of Squares" will be useful later on.

3. Descriptive statistics for standardized scores will be developed in the body of the text.

Standard Scores

It will be recalled that the standard score for an unstandardized measure (raw score) is "the score minus the mean divided by the standard deviation". For case 1 of measure X in Table 1, the standard (z) score is:

$$Z_{x_1} = \frac{X_1 - \bar{X}}{S_x}$$

For any (hypothetical) case, the standard score of an X measure is:

$$Z_{x_i} = \frac{X_i - \bar{X}}{S_x}$$

The same procedure can be applied to Y measures. For case 1:

$$Z_{Y_1} = \frac{Y_1 - \bar{Y}}{S_Y}$$

Similarly, for the ith (hypothetical case), we have:

$$Z_{Y_i} = \frac{Y_i - \bar{Y}}{S_Y}$$

Since a standard score distribution (such as in Table 1) is a distribution of variable measures, we can calculate means, standard deviations, variances, correlations, and so forth, just as we can calculate these statistics for unstandardized measures.

Most students will recall that the mean of z scores is equal to 0 and the standard deviation (and variance) of standardized scores is equal to 1. (The proof of these statements is given in the Appendix[1].)

The mean of X standardized scores is defined as:

$$\bar{z}_x = \frac{\sum_{i=1}^{n_{z_x}} z_{x_i}}{n_{z_x}} = 0$$

Similarly, for Y measures:

$$\bar{z}_y = \frac{\sum_{i=1}^{n_{z_y}} z_{y_i}}{n_{z_y}} = 0$$

The variance of X in z score notation is defined as:

$$s^2_{z_x} = \frac{\sum_{i=1}^{n_{z_x}} (z_{x_i} - \bar{z}_x)^2}{n_{z_x} - 1} = 1$$

[1] The Appendix contains proof of certain concepts or relationships that may be of interest to the reader but are not crucial for the development of the proof in this paper (the theoretical limits of the sample bivariate correlation coefficient).

For the standardized Y measure the variance is defined as:

$$s^2_{z_y} = \frac{\sum_{i=1}^{n_{z_y}} (z_{y_i} - \bar{z}_y)^2}{n_{z_y} - 1} = 1$$

Sum of Squared Standard Scores.

To understand the proof it is necessary to know the result of summing a distribution of squared standard scores. If we square each standard score for the X measure in Table 1 and sum them, we obtain:

$$\sum_{i=1}^{n_{z_x}} z^2_{x_i} = z^2_{x_1} + z^2_{x_2} + \ldots + z^2_{x_n}$$

If we substitute the appropriate means and variances in the right hand side of the expression, we obtain:

$$\sum_{i=1}^{n_{z_x}} z^2_{x_i} = \frac{(X_1 - \bar{X})^2}{s^2_x} + \frac{(X_2 - \bar{X})^2}{s^2_x} + \ldots + \frac{(X_n - \bar{X})^2}{s^2_x}$$

Since the s_x^2 is a constant, we can factor it outside, and write:

$$\sum_{i=1}^{n_{z_x}} z_{x_i}^2 = \frac{1}{s_x^2}\left[(X_1-\bar{X})^2 + (X_2-\bar{X})^2 + \ldots + (X_n-\bar{X})^2\right]$$

Rewriting the right hand side in summation notation, we obtain:

$$\sum_{i=1}^{n_{z_x}} z_{x_i}^2 = \frac{\sum_{i=1}^{n_x} (X_i-\bar{X})^2}{s_x^2}$$

From Table 2, we know that we can substitute the sum of squares term into the numerator on the right hand side. This results in:

$$\sum_{i=1}^{n_{z_x}} z_{x_i}^2 = \frac{(n_x-1)s_x^2}{s_x^2} = n_x-1 = n-1$$

(Recall that $n_x-1 = n-1$).

If we were to work through the same steps for Y, we would obtain:

$$\sum_{i=1}^{n_{z_y}} z_{y_i}^2 = \frac{(n_y-1)s_y^2}{s_y^2} = n_y-1 = n-1$$

(Recall that $n_y-1 = n-1$).

These relationships between squared z scores and sample size are very important for the proof later on. They will be summarized later on for easy reference.

Correlation Formulas

Unstandardized Form

Using the notation and variables in Tables 1 and 2, the unstandardized form correlation for two measures (X and Y) is defined as follows:

$$r_{xy} = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[\frac{\sum_{i=1}^{n_x} (X_i - \bar{X})^2}{n_x - 1}\right]\left[\frac{\sum_{i=1}^{n_y} (Y_i - \bar{Y})^2}{n_y - 1}\right]}}$$

Note that the numerator contains the term n-1 because it is not important or necessary to distinguish between n_x-1 or n_y-1. However, in the denominator it is helpful to distinguish n_x-1 from n_y-1. In any case, all of the sample size terms would be equal to the same numerical value if a correlation coefficient were computed on a set of data (n_x-1 = n_y-1 =n-1).

Standardized Form

The correlation of measure X and measure Y in standard score form is defined as follows:

$$r_{z_x z_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Z_{x_i}-\bar{Z}_x)(Z_{y_i}-\bar{Z}_y)}{\sqrt{\left[\frac{\sum_{i=1}^{n_{z_x}}(Z_{x_i}-\bar{Z}_x)^2}{n_{z_x}-1}\right]\left[\frac{\sum_{i=1}^{n_{z_y}}(Z_{y_i}-Z_y)^2}{n_{z_y}-1}\right]}}$$

It is proven in the Appendix that this correlation formula is equal to:

$$r_{z_x z_y} = \frac{\sum_{i=1}^{n} Z_{x_i} Z_{y_i}}{n-1}$$

If we rearrange this formula by multiplying over the n-1 term, we obtain:

$$(n-1)\,r_{z_x z_y} = \sum_{i=1}^{n} Z_{x_i} Z_{y_i}$$

This relationship will be useful in the proof; it will be restated for easy reference later.

The reader may recall that the same correlation coefficient results when the variables are in raw score form or standard score form. That is:

$$r_{xy} = r_{z_x z_y}$$

This statement is proven in the Appendix. We will restate it prior to the proof for the readers convenience.

Inequalities

Before starting the proof it is necessary to review one further topic: algebraic inequalities. In the proof we are required to manipulate one form of inequality: the form "greater than or equal to" and "less than or equal to". An example will serve as a refresher.

For two variables, say A and B, we can write:

$$A \geq B .$$

which means "A is greater than or equal to B".

Equivalently, we can write:

$$B \leq A$$

which means the same thing: "B is less than or equal to A". For example,

$$3 \geq 1 \text{ or equivalently } 1 \leq 3.$$

All of this may be obvious. What students sometimes forget is what happens when multiplying or dividing by negative quantities. For example, if $3 \geq 1$, and we multiply this inequality by -1, we would obtain:

$$-1[(3) \geq (1)] = -3 \leq -1$$

That is, the inequality sign is reversed when multiplying by a negative number. The same result occurs for more complex expressions. For example:

$$1 - A \geq B$$

Multiplying each side of the inequality by -1, we get:

$$-1[(1-A) \geq (B)] = -(1-A) \leq B \text{ or } A-1 \leq B$$

Example:

$$1 - \tfrac{1}{4} \geq 0$$

Multiplying through by -1, we obtain:

$$-1[(1-\tfrac{1}{4}) \geq (0)] = -(1-\tfrac{1}{4}) \leq 0 = \tfrac{1}{4}-1 \leq 0$$

Summary of Important Concepts:

We have reviewed standard scores(z), correlation formulas and algebraic inequalities. All of these concepts are important to understand the proof that follows. For the readers convenience, we will summarize these concepts for easy reference. This is done in Table 3.

Table 3

Summary of Important Concepts

$$\sum_{i=1}^{n_{z_x}} z_{x_i}^2 = \sum_{i=1}^{n_{z_y}} z_{y_i}^2 = n-1$$

$$r_{xy} = r_{z_x z_y}$$

$$\sum_{i=1}^{n} z_{x_i} z_{y_i} = (n-1) r_{z_x z_y} = (n-1) r_{xy}$$

$$-1[(1-A) \geq (B)] = A-1 \leq B$$

Proof

We are now ready to present the proof. Formally, we want to prove the following statements:

$$r_{xy} \geq -1$$

$$r_{xy} \leq +1$$

Writing each of the statements in one linear form:

$$-1 \leq r_{xy} \leq +1$$

This states the same information as the above two separate statements.

The proof consists of two parts: one part shows the lower limit of r_{xy} (i.e., $r_{xy} \geq -1$), and the second part shows the upper limit of r_{xy} (i.e., $r_{xy} \leq +1$). We will prove the upper limit first.

Proof that $r_{xy} \leq +1$

To prove this limit, we will perform algebraic manipulations on a statement which is mathematically true. That statement is:

$$\sum_{i=1}^{n}(z_{x_i} - z_{y_i})^2 \geq 0$$

In words, the statement means: the sum of squared differences of n standardized value pairs will always be equal to or greater than 0. The reader may refer to Table 1 for clarification. The squared differences are taken for each row (pairs) of Z_{x_i} and Z_{y_i} values starting at Z_{x_1}, Z_{y_1} and continuing down to the last pair of Z's (Z_{x_n}, Z_{y_n}).

Most students readily agree that the squared sum will be greater than 0. But can it ever be exactly equal to 0? Yes, theoretically it can. Referring to Table 1, if one imagines each standardized X and Y measure to have the same numerical value[1], then it is apparent that each difference will be 0; so, the squared value of 0 is also 0. Now, a sum of squared 0's will itself be equal to 0. While it may be unlikely to occur in practice, it is only required that $\sum_{i=1}^{n}(Z_{x_i}-Z_{y_i})^2 \geq 0$ be true in a mathematical sense. Thus, the statement is true. We will expand this squared sum, perform algebraic manipulations and substitutions, and arrive at the proof for the upper limit of the sample correlation coefficient.

The actual steps in the derivation will now be presented. Notes pertaining to the algebra are provided for the readers reference. Refer to Tables 1,2 and 3 as needed. It is suggested that the reader first examine the algebraic statement on the left side of the page. Then read the comment on the right side for explanation. See next page.

[1] That is, within pairs, not all pairs. Example:

Z_{x_i}	Z_{y_i}
1.41	1.41
-.68	-.68
.05	.05
etc.	

That $r_{xy} \leq +1$

Notes

$$\sum_{i=1}^{n}(z_{x_i} - z_{y_i})^2 \geq 0$$

A restatement from before. Squaring each term, we obtain an expansion of the binomial in this form:

$$(A-B)^2 = A^2 + B^2 - 2AB$$

$$\sum_{i=1}^{n}(z^2_{x_i} + z^2_{y_i} - 2z_{x_i}z_{y_i}) \geq 0$$

Distributing the summation operator to each term, and bringing the constant (2) outside the summation sign

$$\sum_{i=1}^{n}z^2_{x_i} + \sum_{i=1}^{n}z^2_{y_i} - 2\sum_{i=1}^{n}z_{x_i}z_{y_i} \geq 0$$

This next step is very important. We will substitute three quantities, all from Table 3. They are:

$$\sum_{i=1}^{n}z^2_{x_i} = n-1$$

$$\sum_{i=1}^{n}z^2_{y_i} = n-1$$

$$\sum_{i=1}^{n}z_{x_i}z_{y_i} = (n-1)r_{z_xz_y} = (n-1)r_{xy}$$

Making these three substitutions

$$(n-1) + (n-1) - 2(n-1)r_{xy} \geq 0$$

Collecting the like terms of (n-1)

$$2(n-1) - 2(n-1)r_{xy} \geq 0$$

Factoring the 2(n-1) term

$$2(n-1)[1-r_{xy}] \geq 0$$

Dividing each side of the inequality by 2(n-1) which <u>does not</u> change the inequality sign as 2(n-1) is always positive because n must always be ≥ 2

$$\frac{2(n-1)[1-r_{xy}]}{2(n-1)} \geq \frac{0}{2(n-1)}$$

$$(1-r_{xy}) \geq 0$$

Here we make use of multiplying an inequality by a negative number. Let us multiply each side of the inequality by -1 (see Table 3) which reverses the inequality sign and reverses the $1-r_{xy}$

$$-1[(1-r_{xy}) \geq 0] = r_{xy} -1 \leq 0$$

Now, add +1 to each side

$$r_{xy} - 1 + 1 \leq 0 + 1$$

This gives us

$$r_{xy} \leq +1$$

END OF PROOF FOR UPPER LIMIT.

Proof that $r_{xy} \geq -1$

Part two of the proof will be much simpler because the structure of this part of the proof is very much like the first part. We will follow the same basic steps. We start out with a statement that is mathematically true, namely:

$$\sum_{i=1}^{n}(Z_{x_i} + Z_{y_i})^2 \geq 0$$

Again this statement is true in a mathematical sense even through the "equals 0" aspect is very unlikely to occur in statistical practice.

The development of the proof with appropriate notes begins on the next page.

That $\frac{r_{xy} \geq -1}{2}$

19

$$\sum_{i=1}^{n}(Z_{x_i} + Z_{y_i})^2 \geq 0$$

Step 1 restated. Squaring each term results in a binomial expansion in this form:

$$(A+B)^2 = A^2 + B^2 + 2AB$$

$$\sum_{i=1}^{n}(Z_{x_i}^2 + Z_{y_i}^2 + 2Z_{x_i}Z_{y_i}) \geq 0$$

Distributing the summation operator and bringing out the 2

$$\sum_{i=1}^{n}Z_{x_i}^2 + \sum_{i=1}^{n}Z_{y_i}^2 + 2\sum_{i=1}^{n}Z_{x_i}Z_{y_i} \geq 0$$

Making the same three substitutions as in part one, we obtain

$$(n-1) + (n-1) + 2(n-1)r_{xy} \geq 0$$

Adding like terms and factoring

$$2(n-1)[I + r_{xy}] \geq 0$$

Dividing each side by 2(n-1)

$$1 + r_{xy} \geq 0$$

Adding -1 to each side

$$1. + r_{xy} -1 \geq 0 - 1$$

Simplifying

$$r_{xy} \geq -1$$

END OF PROOF FOR LOWER LIMIT

We have just proven that $-1 \leq r_{xy} \leq +1$. See the Appendix for additional proofs of related material.

APPENDIX

Selected Proofs

1. That the mean of standard scores is equal to 0.

We will start with the definition of the mean of z scores for the X measure.

$$\bar{z}_x = \frac{\sum_{i=1}^{n_{z_x}} z_{x_i}}{n_{z_x}}$$

Expanding the right side:

$$\bar{z}_x = \frac{1}{n_{z_x}}\left[\frac{(X_1-\bar{X})}{S_x} + \frac{(X_2-\bar{X})}{S_x} + \dots + \frac{(X_n-\bar{X})}{S_x}\right]$$

Factoring the constant, S_x, outside and rewriting the sum of deviations in summation notation:

$$\bar{z}_x = \frac{1}{n_{z_x}}\,\frac{1}{S_x}\sum_{i=1}^{n_x}(X_i-\bar{X})$$

Distributing the summation sign inside the parentheses :

$$\bar{z}_x = \frac{1}{n_{z_x}} \frac{1}{s_x} \left[\sum_{i=1}^{n_x} X_i - \sum_{i=1}^{n_x} \bar{X} \right]$$

Since $\sum_{i=1}^{n_x} X_i = n_x \bar{X}$ and the sum of the constant, $\sum_{i=1}^{n_x} \bar{X}$, is equal[1] to $n_x \bar{X}$,

$$\bar{z}_x = \frac{1}{n_{z_x}} \frac{1}{s_x} (n_x \bar{X} - n_x \bar{X}) = 0$$

Thus the mean of z_x scores is equal to 0. Similar reasoning for the Y measure will produce the same result, namely:

$$\bar{z}_y = \frac{\sum_{i=1}^{n_{z_y}} z_{y_i}}{n_{z_y}} = \frac{1}{n_{z_y}} \frac{1}{s_y} (n_y \bar{Y} - n_y \bar{Y}) = 0$$

Therefore, variables in standardized form have mean equal to 0.

[1] Recall that when taking the sum of a constant (say C), we have:

$$\sum_{i=1}^{n} C = C + C + C + \ldots + C = nC$$

That is, the sum of a constant is equal to the constant times the number of terms added (in this case, n).

2. That the variance and standard deviation of standard scores is equal to 1.

By definition, the variance for X measures in standard score form is:

$$s^2_{z_x} = \frac{\sum_{i=1}^{n_{z_x}} (Z_{x_i} - \bar{Z}_x)^2}{n_{z_x} - 1}$$

Since we know that $\bar{Z}_x = 0$, we now have:

$$s^2_{z_x} = \frac{\sum_{i=1}^{n_{z_x}} (Z_{x_i})^2}{n_{z_x} - 1}$$

If we rewrite Z_{x_i} in terms of unstandardized mean and standard deviation:

$$s^2_{z_x} = \frac{1}{n_{z_x} - 1} \sum_{i=1}^{n_x} \left[\frac{(X_i - \bar{X})}{s_x}\right]^2$$

Rearranging terms:

$$s^2_{z_x} = \frac{1}{n_{z_x} - 1} \frac{1}{s^2_x} \sum_{i=1}^{n_x} (X_i - \bar{X})^2$$

From Table 3, we can substitute into the numerator of $S^2_{z_x}$ the "sum of squares" for the X measure. This results in:

$$S^2_{z_x} = \frac{1}{n_{z_x}-1} \frac{1}{S^2_x} (n_x-1)(S^2_x)$$

Since $n_x = n_{z_x}$ we can cancel terms, leaving:

$$S^2_{z_x} = 1$$

Similar reasoning for Y standardized measures will produce, as the next to the last step in the derivation:

$$S^2_{z_y} = \frac{1}{n_{z_y}-1} \frac{1}{S^2_y} (n_y-1)(S^2_y)$$

Since $n_y = n_{z_y}$

$$S^2_{z_y} = 1$$

In each case, the standard deviation for appropriate variance terms is simply the square root of 1. That is:

$$\sqrt{S^2_{z_x}} = S_{z_x} = \sqrt{1} = 1 \text{ and } \sqrt{S^2_{z_y}} = S_{z_y} = \sqrt{1} = 1$$

Thus, the variance and standard deviation of z scores is equal to 1.

3. That $r_{xy} = r_{z_x z_y}$

We want to show that when measures X and Y are converted to standard scores and correlated, the resulting correlation is the same as the correlation between the unstandardized (raw) measures of X and Y.Let us first rewrite the correlation formula for z scores:

$$r_{z_x z_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Z_{x_i}-\bar{Z}_x)(Z_{y_i}-\bar{Z}_y)}{\sqrt{\left[\frac{\sum_{i=1}^{n_{z_x}}(Z_{x_i}-\bar{Z}_x)^2}{n_{z_x}-1}\right]\left[\frac{\sum_{i=1}^{n_{z_y}}(Z_{y_i}-\bar{Z}_y)^2}{n_{z_y}-1}\right]}}$$

Since $\bar{Z}_x = \bar{Z}_y = 0$, we can simplify to get:

$$r_{z_x z_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Z_{x_i})(Z_{y_i})}{\sqrt{\left[\frac{\sum_{i=1}^{n_{z_x}}(Z_{x_i})^2}{n_{z_x}-1}\right]\left[\frac{\sum_{i=1}^{n_{z_y}}(Z_{y_i})^2}{n_{z_y}-1}\right]}}$$

In the denominator, we recognize that $\sum_{i=1}^{n_{z_x}}(Z_{x_i})^2 = n_{z_x}-1$ and

$\sum_{i=1}^{n_{z_y}}(Z_{y_i})^2 = n_{z_y}-1$. Substituting these values, we obtain:

$$r_{z_x z_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Z_{x_i})(Z_{y_i})}{\sqrt{\left[\frac{n_{z_x}-1}{n_{z_x}-1}\right]\left[\frac{n_{z_y}-1}{n_{z_y}-1}\right]}}$$

The denominator cancels out completely leaving:

$$r_{z_x z_y} = \frac{1}{n-1}\sum_{i=1}^{n} Z_{x_i} Z_{y_i}$$

(Recall that this relationship was used in the proof for the limits of r_{xy}).

Now, expanding the z score terms:

$$r_{z_x z_y} = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{(X_i-\bar{X})(Y_i-\bar{Y})}{(S_x)(S_y)}\right]$$

This is identical to :

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{(S_x)(S_x)}$$

Recognizing that $\sqrt{S_x^2} = S_x$ and $\sqrt{S_y^2} = S_y$, we can write:

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{(S_x^2)(S_y^2)}}$$

Rewriting the denominator of the variance product term in raw score terms (see Table 2):

$$r_{z_x z_y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\left[\frac{\sum_{i=1}^{n_x}(X_i-\bar{X})^2}{n_x-1}\right]\left[\frac{\sum_{i=1}^{n_y}(Y_i-\bar{Y})^2}{n_y-1}\right]}}$$

This is precisely the form for r_{xy} that was defined earlier in the paper. Therefore, the correlation between measures in raw score and z score forms is identical.

References

Glass, Gene V and Stanley, Julian C. Statistical Methods in Education and Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

O'Brien, Francis J., Jr. A proof that t^2 and F are identical: the general case. ERIC Clearinghouse on Science, Mathematics and Environmental Education, Ohio State University, April, 1982. (Note: the ED number for this document was not available at the time of this publication).